14/04/2017

Rapport de stage

De Nihed Ghédira

M.A. Brahem

Sous la direction d'A. Belaïd

LORIA

Analyse et annotation des images de transactions du Musée de Musique (Paris)

Sommaire

Remerciements

« Nous remercions le LORIA pour nous avoir accueillis durant ces 3 mois.

Nous tenons à remercier tout particulièrement Mr Belaïd le responsable de l'équipe READ, qui nous a accordé sa confiance et nous a accordé des missions valorisante durant ce stage, et Mr Karpinski qui a su trouver du temps pour nous aider.

Faire un stage au sein du LORIA a été un grand plaisir. Nous avons pu découvrir le monde des chercheurs grâce à vous. »

Introduction

Motivations et objectives

Le sujet général de cette étude consiste à réaliser une étude de faisabilité sur la reconnaissance de mots clés dans des images numérisées représentant des transactions de vente de violons, appartenant au Musée de la Musique à la Villette (Paris). Dans ce stage, nous nous sommes focalisées sur l'annotation du contenu des images ainsi que sur quelques prétraitements d'images afin de permettre l'extraction des mots.

Dans la suite, nous allons évoquer le cahier des charges du Musée, puis nous détaillerons le travail réalisé au cours de ce stage.

Cahier des charges du Musée

« Le document proposé au test est un registre provenant du fonds d'archives de l'atelier de lutherie parisien Gand, Bernardel, Caressa et Français.

Ce fonds couvre un siècle et demi d'histoire, de 1816 à 1944. Cet atelier de lutherie fondé par Nicolas Lupot à Paris en 1796 a connu un destin exceptionnel en raison de :

- la figure de son fondateur, Nicolas Lupot appelé le « Stradivarius français » et celles de ses successeurs,
- l'importance et le prestige de sa clientèle,
- l'excellence dans la facture d'instruments,
- l'expertise,
- la restauration
- le commerce d'instruments anciens.

Cet atelier est probablement le plus important atelier de lutherie français, voire mondial, du XIXe siècle par sa longévité, par son prestige, par sa clientèle internationale.

Les documents qui composent ce fonds, sont principalement des registres de réparations et de ventes d'instruments. Nous avons sélectionné ce registre (E.981.8.38), qui concerne la vente d'instruments neufs et anciens entre 1840 et 1902 parce qu'il présente la meilleure (la moins pire) des structures de données de l'ensemble des registres.

Le document et tout le fonds en général, est une source importante pour l'histoire des violons prestigieux comme les stradivarius, guarnérius, amati… et pour la mention de grands musiciens ou collectionneurs. On y trouve la trace des ventes, donc de la valeur attribuée aux instruments, ainsi que les mentions de restaurations et transformations apportées à ces instruments.

En revanche, la recherche n'est pas facile dans ces archives car :

- le fonds représente 11.000 vues avec des documents qui n'ont pas tous la même structure
- le texte est manuscrit et de différentes mains
- les instruments ne sont pas répertoriés dans des index.

La recherche par reconnaissance de texte permettrait donc de faire une recherche « transversale ».

Du point de vue de l'historien, les priorités seraient :

- retrouver les noms des luthiers prestigieux
- retrouver les noms de quelques propriétaires importants
- connaître la valeur financière des instruments lors des transactions

Quelques-uns des noms recherchés (Les mots et champs clés)

- Stradivari - Stradivarius
- Amati
- Guarneri - Guarnerius
- Guadagnini
- Lupot
- Tourte

Pour les documents qui n'ont pas de répertoire (comme le registre E.981.8.38 proposé en test), il faudrait ajouter :

- Conservatoire
- Clapisson
- Hamma
- Wurlitzer
- Tolbecque
- Chanot
- Chardon
- Besson
- Enesco
- Opéra »

La mission du Musée est d'acquérir, de préserver et de valoriser ses collections afin de contribuer à la sauvegarde du patrimoine naturel, culturel et scientifique. Ses collections constituent un important patrimoine public, occupent une position particulière au regard de la loi et jouissent de la protection du droit international.

À cette mission d'intérêt, il faut qu'on annote les documents scannés d'une façon précise qui n'introduit aucune perte au niveau d'écriture, date et numéros c'est-à-dire un véritable document de vérité.

Description de la base des registres

Le contenu

La base de registres concerne les livres de type “registre de ventes/réparation”. Elle contient une centaine d’images en .jpg, scannées à 300 ppp. Chaque image décrit deux pages successives. La Figure 1 donne un exemple de ces pages. Chaque page contient des transactions de vente. L’année est indiquée à gauche pour un ensemble de transactions. Ensuite, on trouve pour chaque transaction, un numéro d’ordre, le texte de la transaction, suivi par un prix estimé et le prix de vente. En cas d’annulation de la vente, le texte est barré et corrigé (reprise pour une deuxième vente), ce qui le rend parfois illisible. Le texte contient essentiellement des noms de personnes et d’instruments qu’il faudra extraire.

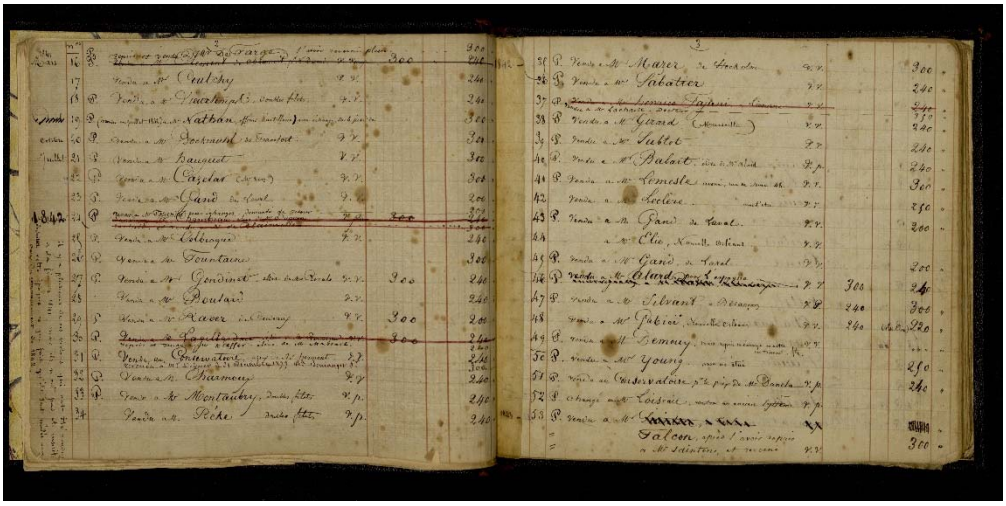


Figure 1: image d'une double page du registre

Les images numérisées

D’un point de vue image, ces images sont complexes pour les raisons suivantes :

- Le papier est ancien occasionnant des tâches et des différences de contrastes.
- L’écriture est cursive et attachée créant des ligatures entre les mots et les digits et rendant la segmentation complexe.
- La pliure des pages au centre crée une déformation (un bombage) qui altère l’alignement des lignes de texte. De plus, la page de gauche est orientée vers la gauche, et celle de droite, vers la droite. Ceci se voit sur les lignes graphiques, séparant les colonnes, qui sont inclinées.
- Un cadre noir entoure souvent le texte et a une configuration différente pour chaque image.
- Enfin, les documents étant multi-scripteurs, l’écriture change ainsi que la structure du contenu

La structure

On peut dégager dans une page deux types de structures :

- Structure physique
    - Elle décrit les éléments du document tels qu’ils apparaissent disposés dans la page (voir Figure 2). On découpe la page horizontalement en colonnes. Chaque colonne contient un type d’information particulier sur la transaction. Ainsi, on peut trouver une colonne de date, une colonne de numéro d’ordre, etc.

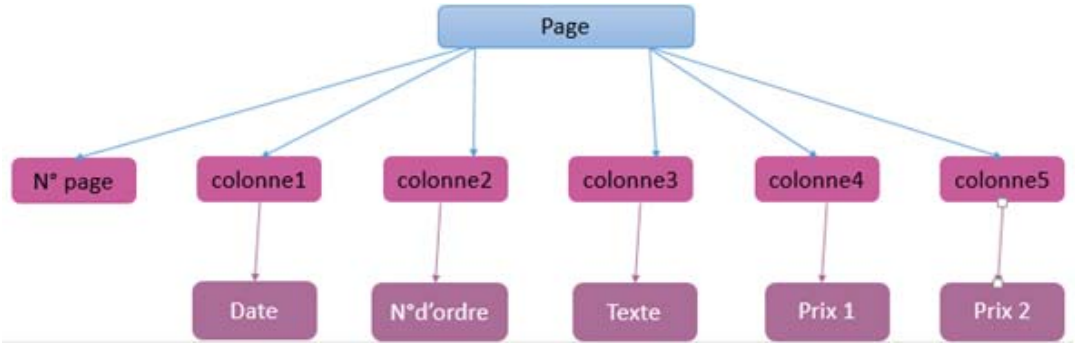


Figure 2: structure physique de la page

- Structure logique
  - La structure logique décrit le contenu tel qu'on doit le lire. Ainsi, une page est décrite par un numéro de page et des transactions. La transaction composée de trois parties : date, N° d'ordre et corps. Le corps contient le texte de la transaction ainsi que les prix estimé et réel (voir Figure 3).

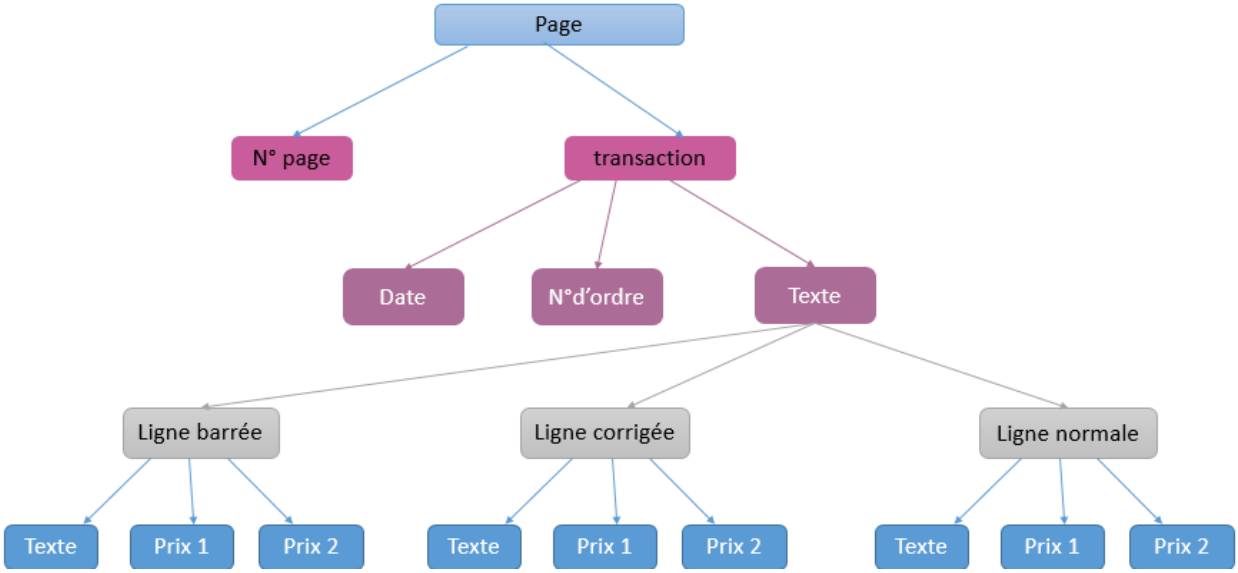


Figure 3: structure logique de la page

Prétraitement d'images

Le prétraitement consiste à nettoyer les images de leur bruit, enlever le cadre noir qui les entoure, binariser les images et détecter les colonnes graphiques. Nous avons utilisé quelques méthodes existantes de l'état de l'art pour réaliser ces opérations.

Suppression du cadre

Pour la suppression du cadre on a implémenté le modèle de contour actif, également appelé « snake » qui était introduit par Kass et Witkin en 1987. C'est une technique d'extraction de données, la reconnaissance de formes et la détection de bord.

Principe :

C'est une méthode dont le principe consiste à placer dans l'image au voisinage de la forme à détecter un contour initial qui sera ensuite déformé sous l'action de plusieurs forces :

- L’énergie externe permettant de régulariser le contour
- L’énergie potentielle reliée à l’image
- L’énergie interne reliée aux contraintes particulières que l’on peut ajouter.

1) L’énergie externe :
Correspond à l'impact du contour sur l'image. Pour la calculer, il faut considérer l'opposé de la valeur de son gradient (ou de toute autre représentation mettant en jeu les contours à épouser) en chaque point du contour. Cette énergie externe doit théoriquement être minimale si le contour épouse parfaitement la forme à extraire.

2) L’énergie potentielle
Liée à l’image, elle représente les éléments sur l’image vers lesquels on veut attirer le Snake.

3) L’énergie interne :

Ne dépend pas de l'image ni de la forme à détourer, elle ne dépend que des points du contour. Elle regroupe des notions comme la courbure du contour ou la régularité d'espacement des points. Ces énergies vont permettre au contour actif d’évoluer pour rechercher la position d’énergie minimale, représentant le contour recherché. La Figure 4 montre comment ces énergies évoluent autour de la bordure de la forme blanche.

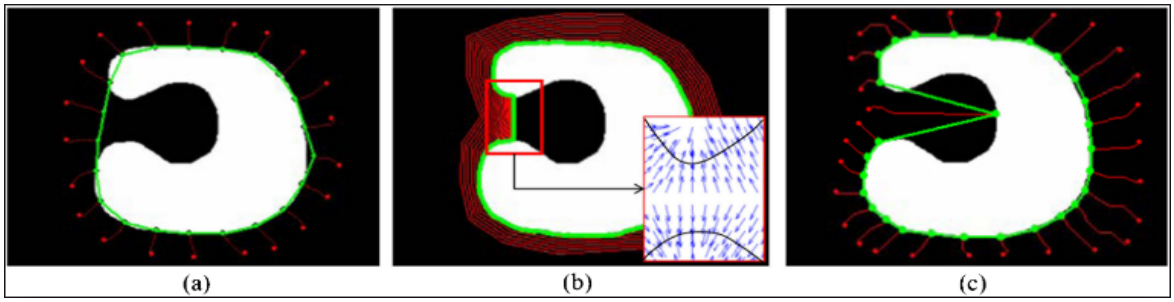


Figure 4: image de l'évolution de l'énergie autour de la bordure

L’actif contour peut être compris comme un cas particulier de la technique générale d'adaptation d’une bordure déformable, c’est-à-dire, le contour varie d’une image à une autre puisque les documents sont scannés.

Γ0 est le contour initial qui peut être défini manuellement, par exemple, et v est la vitesse d'évolution de la courbe. La figure 7 illustre cette évolution.

L’algorithme va tenter de trouver un meilleur positionnement pour le contour pour minimiser les dérives par rapport aux contraintes utilisées. Il s'arrêtera lorsqu'il ne sera plus possible d'améliorer le positionnement ou simplement quand le nombre maximum d'itérations aura été atteint. On utilise les notions d'énergies interne et externe pour caractériser respectivement la forme du contour et tous les éléments qui lui sont propres, et le positionnement du contour sur l'image en tenant compte des lignes de gradient.

La Figure 5 montre comment l’actif contour peut fonctionner sur un cadre d’image de registre.

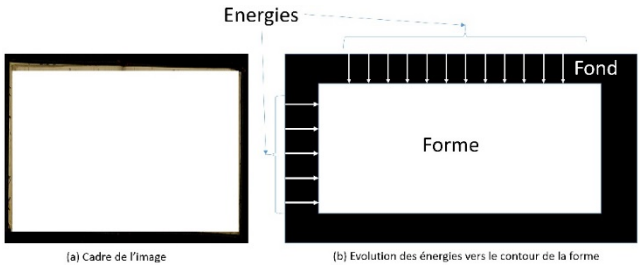


Figure 5: Cadre de l'image des registres

Le Snake est initialisé sur le bord externe de l'image. Il évolue progressivement vers la forme (intérieur de l'image) en faisant des itérations au cours desquelles il calcule le gradient. Comme la bande noire contient un gradient nul, le Snake évolue naturellement vers le contour intérieur de la bande et donc, vers la forme. Même si la forme (ici le texte) est un peu attaché à la bordure, les attaches vont être coupées par manque d'énergie forte au niveau de ces attaches.

La Figure 6 donne le code Matlab de l'Active contour. Le contour initial est le masque de la convolution permettant d'extraire le gradient.

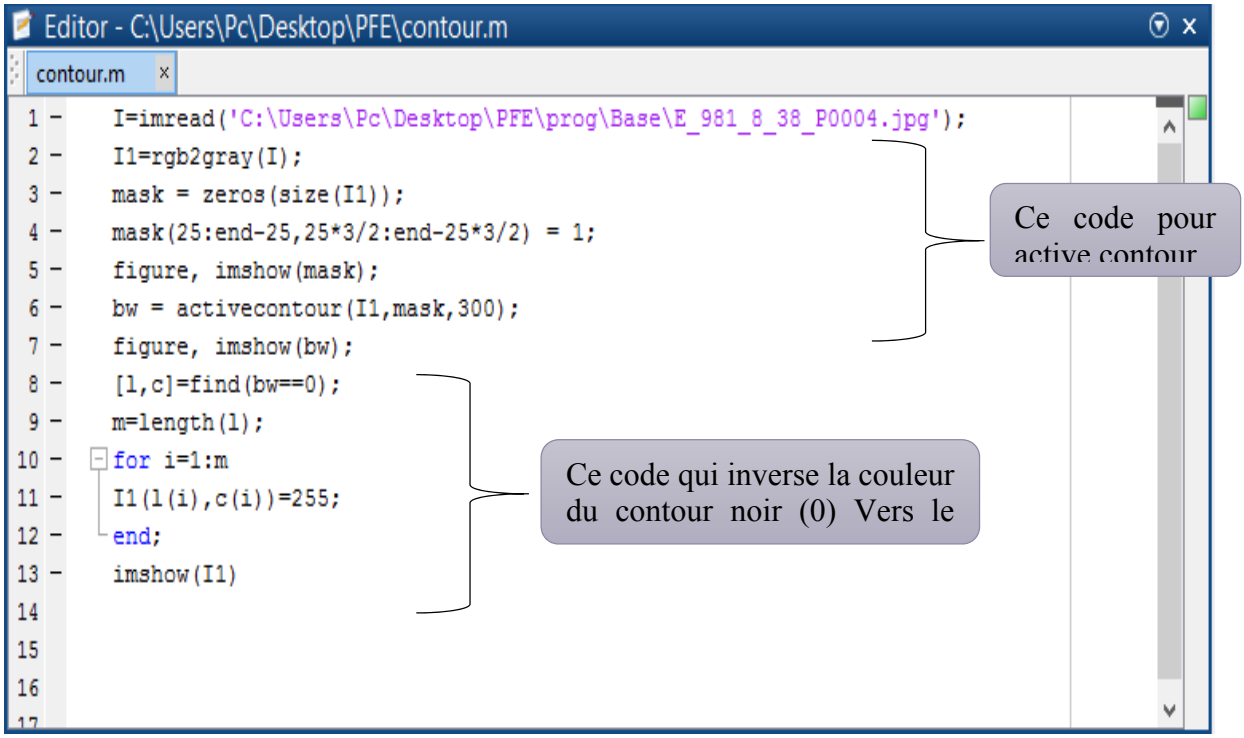


Figure 6: code Matlab de l'Active Contour

La Figure 7 donne l'Active contour au début. La Figure 8 présente le résultat de l'extraction. La Figure 9 donne le résultat de la suppression de la bande dans l'image d'origine.

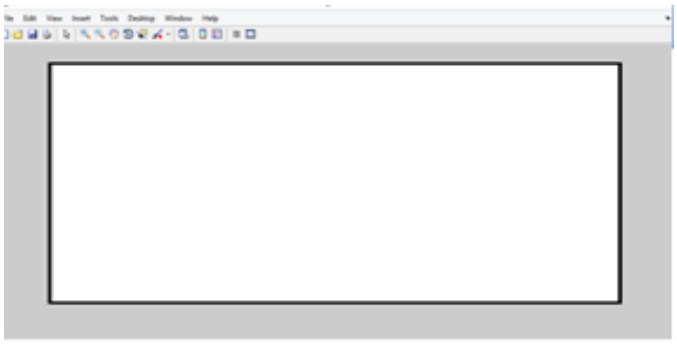


Figure 7: le résultat d'imshow1 qui est le contour initial

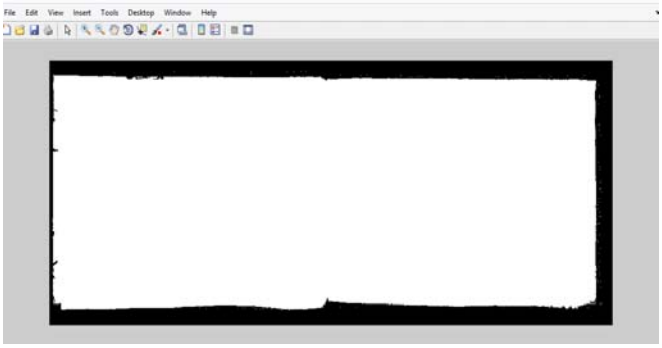


Figure 8: le résultat d'imshow2 qui donne la bordure détectée

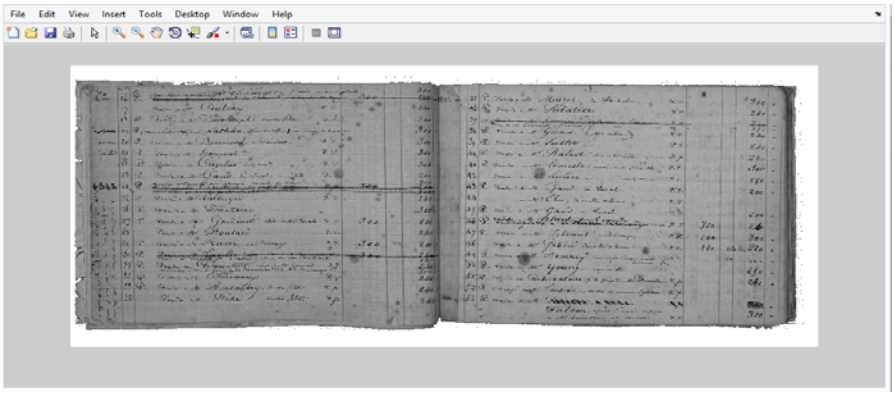


Figure 9: le résultat d'imshow3 qui indique l'image sans bordure

Binarisation

L'objectif de la binarisation est de simplifier l'image couleur en la rendant binaire, et aussi de la débarrasser des tâches qui l'altèrent, sans pour autant détruire les traits de l'écriture.

On a utilisé deux algorithmes existants dans la littérature :

- Cohen et al.i
- Kligler Nati et Ayellet Talii

Méthode Cohen et al.

Elle est basée sur la notion de « component tree » et est ajustée pour extraire correctement les lignes. Elle est orientée : extraction de lignes. L'algorithme applique d'abord un smearing (étalement), puis extrait les lignes. La méthode est proche de celle de Shi et al. qui convertit l'image en carte de points de connectivité où la valeur de chaque pixel est définie par l'intensité cumulée à l'intérieur d'une fenêtre d'une certaine dimension. Ensuite l'image est binarisée en utilisant un seuil.

Figure 10: Binarisation par l'algorithme de Cohen et al.

Méthode de Kligler Nati et Ayellet Tal

Ces auteurs ont participé et gagné la compétition ICFHR 2016. Le processus de binarisation s'appuie sur une technique de prétraitement très performante. Comme le disent les auteurs : « L'image est considérée comme un ensemble de points 3D (X, Y, intensité). Cet ensemble est transformé linéairement à partir de l'espace 3D Euclidien sur une surface sphérique. Lors de l'application de la transformation, on peut montrer que les concavités sur la surface de la sphère correspondent au texte dans l'image originale. Ensuite, afin de détecter ces concavités, ils utilisent l'opérateur d'élimination des points cachés (HPR), tels que décrit dans (S. Katz and A. Tal, "Direct Visibility of Point Sets," SIGGRAPH, vol. 26, no. 3, 2007.) ».

Figure 11: Image binarisée par l'algorithme de Kligler et Tal

En comparant les deux techniques, on constate que l'algorithme de Kligler et Tal enlève en même temps le cadre, l'écriture est beaucoup plus lisible que dans celle de Cohen et al. et que ce dernier enlève des points de l'écriture, ce qui peut être dommageable pour elle.

Extraction des colonnes

Nous rappelons que les colonnes représentent les différentes rubriques qui constituent les transactions. Elles sont séparées par des lignes graphiques verticales. Nous cherchons à extraire ces lignes graphiques. Nous avons utilisé quatre méthodes :

- Méthode de couleur
- Méthode de Hough
- Méthode de projection
- Méthode de pattern

Méthode de couleur

Improfile (méthode Matlab) calcule les valeurs d'intensité le long d'une ligne dans une image.
Elle sélectionne des points également espacés le long du chemin spécifié, puis utilise l'interpolation pour trouver la valeur d'intensité pour chaque point. Improfile fonctionne avec des images d'intensité de gris et des images RVB. La Figure 12 donne le code source de la méthode.

```
function Profil_Callback(hObject, eventdata, handles)
% hObject    handle to Profil (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
axes(handles.axesorig);
img=handles.tochanged;
[rows columns numberOfColorChannels] = size(img);
if numberOfColorChannels > 1
    % Color
    img = rgb2gray(img);
else
    % nvg or binary
    maxValue = max(img(:));
    if maxValue==1
        img=im2uint8(img);
    end
end
improfile
```

Figure 12: Le code de l'affichage et de profil de l'image

La Figure 13 montre les pics de l'histogramme du profil qui correspondent aux lignes graphiques.

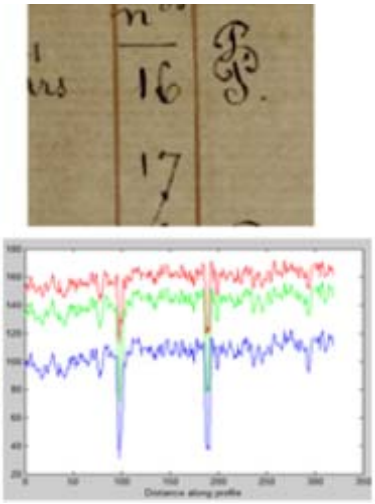


Figure 13: Résultat de la fonction improfile

Méthode de Hough

La transformée d'Hough est une technique de reconnaissance de formes inventée en 1962 par Paul Hough et breveté par IBM. Elle permet de détecter des objets bien précis dans les images:

- Des droites
- Des cercles, des rectangles, des ellipses…

Cette méthode a été utilisée par Likforman-Sulemetal. [Likforman-S1995] pour extraire les lignes de texte dans des documents manuscrits et a été également utilisée pour extraire les lignes dans des documents manuscrits de différents types (lettres, notes, etc.)[Malleron2009]

- Haut du formulaire

  Pour appliquer la transformée d'Hough à une image de largeur L et de hauteur H, il faut créer un espace d'Hough.

- Il faut discrétiser l'espace, en abscisse de -π/2 à π/2, en ordonnée de -d'à +c'ou d'est la taille de la diagonale de l'image).
- Créer un accumulateur, et initialiser tous ses cases à 0.
- Parcourir les pixels des images, on opère de la manière suivante:
  - On fixe θ et on calcule r=x cos(θ) + y Sin(θ)
  - Ajout de vote pour [r] [θ]
  - Incrémentation de la valeur de la case correspondant Algorithme
  - Bas du formulaire

On s'est inspiré de l'algorithme de Hough en faisant quelques modifications dans les paramètres de précision de code pour détecter les colonnes. Puis la confirmation se fait avec la projection. La transformée de Hough est un outil efficace pour détecter les droites dans une image. Il existe d'autres transformées de Hough, dites transformées de Hough généralisées pour extraire d'autres formes.

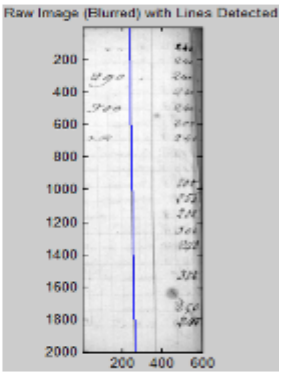
Elle est utilisée dans plusieurs applications :

- Détection des routes dans les images prises par satellite
- Lecture de code-barres…

Dans la Figure 14, on remarque que dans l'exemple(a) le paramétrage est moins précis que l'exemple (c) ce qui implique la détection d'une seule ligne graphique (b), or dans le deuxième exemple la précision est maximale donc la détection est correcte

```
I=img;
 fltr4img = [1 2 3 2 1; 2 3 4 3 2; 3 4 6 4 3; 2 3 4 3 2; 1 2 3 2 1];
 fltr4img = fltr4img / sum(fltr4img(:));
 imgfltrd = filter2( fltr4img , I );
 tic;
 [accum, axis_rho, axis_theta, lineprm, lineseg] = ...
     Hough_Grd(imgfltrd,  10, 0.55);
 toc;
  figure(2); imagesc(imgfltrd); colormap('gray'); axis image;
 DrawLines_Polar(size(imgfltrd), lineprm);
 title('Raw Image (Blurred) with Lines Detected');
```

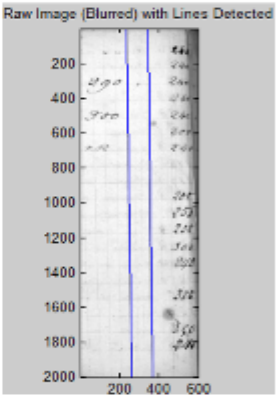
(a)



(b)

```
I=img;
 fltr4img = [1 2 3 2 1; 2 3 4 3 2; 3 4 6 4 3; 2 3 4 3 2; 1 2 3 2 1];
 fltr4img = fltr4img / sum(fltr4img(:));
 imgfltrd = filter2( fltr4img , I );
 tic;
 [accum, axis_rho, axis_theta, lineprm, lineseg] = ...
     Hough_Grd(imgfltrd,  6, 0.45);
 toc;
  figure(2); imagesc(imgfltrd); colormap('gray'); axis image;
 DrawLines_Polar(size(imgfltrd), lineprm);
 title('Raw Image (Blurred) with Lines Detected');
```

(c)



(d)

Figure 14:deux exemples de paramétrage de Hough

Méthode de projection

La méthode la plus générique pour la détection des lignes graphiques est la projection verticale. Dans notre cas, elle est basée simultanément sur la projection verticale et la détection des pics.

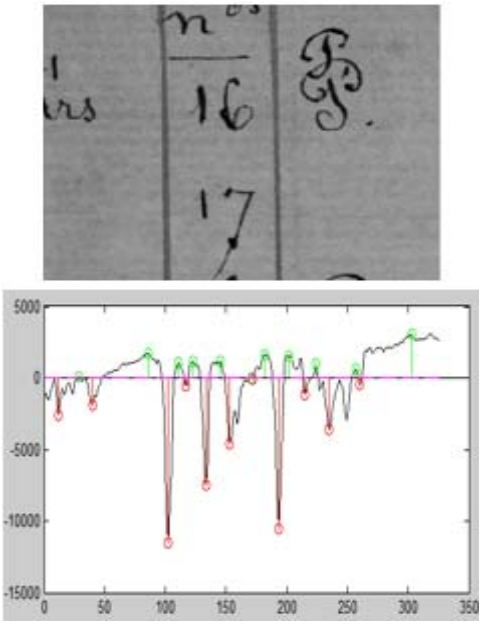


Figure 15:Résultat de la projection verticale avec la détection des pics

La présence de plusieurs pics et plusieurs vallées confirme la présence des lignes graphiques verticales.

Dans le contexte du traitement d'image, un profil de projection verticale est nécessaire pour identifier ou détecter les lignes d'un texte. Chaque pic correspond à une ligne graphique.

La détection de pics est une procédure commune en matière d'analyse du signal, ou de trouver des maximum locaux, des valeurs plus grandes que les points de données adjacents, dans un signal bruité.

En effet, cette technique échoue à détecter les lignes graphiques puisque ils sont inclinées.

Méthode de pattern

Après une première réflexion sur l'extraction des colonnes, on a pensé à construire un modèle à partir de la position des colonnes dans chaque image. Pour savoir si les colonnes sont structurées d'une façon répétée, On a suivi les étapes le suivant :

- Ouvrir l'image avec un éditeur d'image.
- Positionner le curseur sur le pixel de l'extrémité d'une colonne
- Noter sa position sur l'axe des abscisses comme indique le rectangle rouge dans la Figure 16.

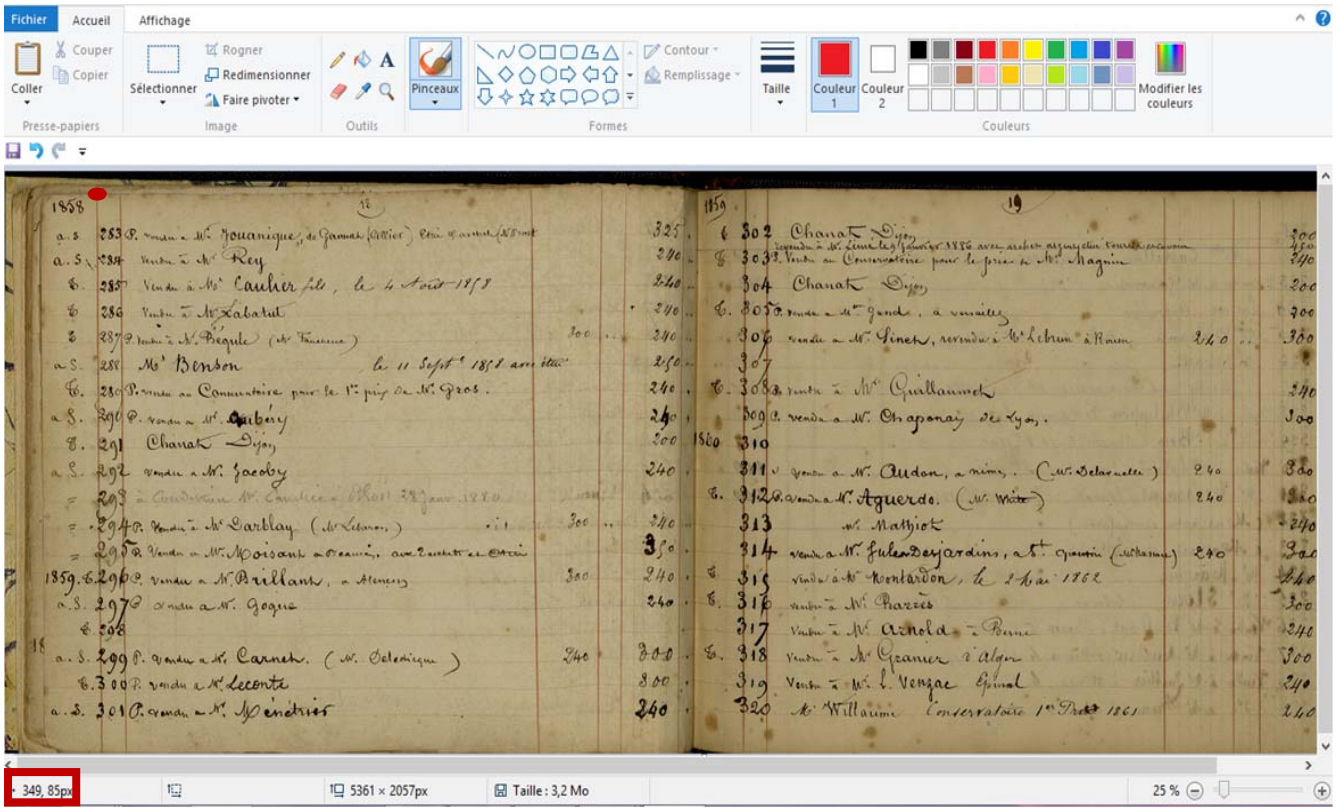


Figure 16:Position du curseur dans l'image

- Faire la différence entre l'origine et l'extrémité de chaque colonne

Les colonnes	origine	Extrémité	Différence
1	345	385	40
	441	477	36
2	1997	2021	24
	2325	2345	20
3	2449	2473	24
	2681	2705	24
4	2937	2913	24
	3025	3001	24
5	4565	4533	32
	4825	4801	24
6	4929	4897	68
	5165	5137	32
7	5265	5237	32

Figure 17: Tableau comparatif entre les deux extrémités

- Refaire sur plusieurs images.

		Page1									
Intitulé		Colonne 1		Colonne 2		Colonne 3		Colonne 4		Colonne 5		Colonne 6	
E_981_8_38_P0004.jpg	Origine	345	441	96	1997	1556	2325	328	2449	124	2681	232
	Extrimité	385	447	62	2021	1574	2345	324	2473	128	2705	232
	Déf	40	36	-4	24	-12	20	-4	24	4	24	0
E_981_8_38_P0011.jpg	Origine	353	441	88	2017	1576	2285	268	2381	96	2601	220
	Extrimité	373	465	92	2049	1584	2281	232	2365	84	2585	220
	Déf	20	24	4	32	8	4	-28	16	12	16	0
E_981_8_38_P0050.jpg	Origine	521	609	88	2141	1532	2405	264	2501	96	2601	100
	Extrimité	545	633	88	2155	1522	2419	264	2517	98	2585	68
	Déf	24	24	0	14	-10	11	-3	16	5	16	0
E_981_8_38_P0070.jpg	Origine	521	602	81	2133	1531	2397	264	2497	100	2837	340
	Extrimité	557	645	88	2149	1504	2413	264	2513	100	2849	336
	Déf	36	43	7	16	-27	16	0	16	0	12	-4
E_981_8_38_P0088.jpg	Origine	681	773	92	2281	1508	2541	260	2645	104	2861	216
	Extrimité	617	705	88	2237	1532	2505	268	2597	92	2833	236
	Déf	64	68	4	44	-24	36	-8	48	12	12	-36
E_981_8_38_P0094.jpg	Origine	513	605	92	2129	1524	2397	268	2497	100	2729	232
	Extrimité	505	597	92	2113	1516	2385	272	2481	96	2713	232
	Déf	8	8	0	11	3	12	1	16	4	16	0

Figure 18: Tableau comparatif entre les images

Le tableau ci-dessus indique l'inclinaison d'une colonne (Déf), la distance entre deux origines colonnes dans la même page (cellules jaunes) ainsi que la distance entre deux extrémités (cellules vert).

On constate alors que l'arrangement des colonnes et différent d'une page à une autre ce qui nous empêche de créer un modèle et appliquer cette méthode.

Annotation

Objectif

Après l'étude de la forme du livre, l'annotation des images est nécessaire pour :

- Avoir de la métadonnée qui résume les documents images. On l'utilisera après pour accéder aux données de registre.
- La Création d'une base de données (**dataset**) pour l'apprentissage et l'évaluation d'un système de reconnaissance optique de caractères (**OCR**).

Présentation de GEDI

GEDI (Groundtruthing **E**nvironement for **D**ocument **I**mages) a été conçue par LAMP[iii]. C'est un outil qui assiste à construire un ensemble de documents images annotées« Ground truth ». Le concept général d'annotation de GEDI est d'interpréter le contenu d'un document image comme un ensemble de zones, et chaque zone peut avoir un ensemble d'attributs. Chaque attribut et construit alors d'un pair Nom/Valeur. Le design de GEDI aide les utilisateurs à spécifier leurs propres attributs et de personnaliser leurs utilités. Des outils sont utilisés dans l'interface pour prévoir une annotation simplifiée.

Dans le reste de cette partie on va présenter l'interface et ses fonctionnalités.

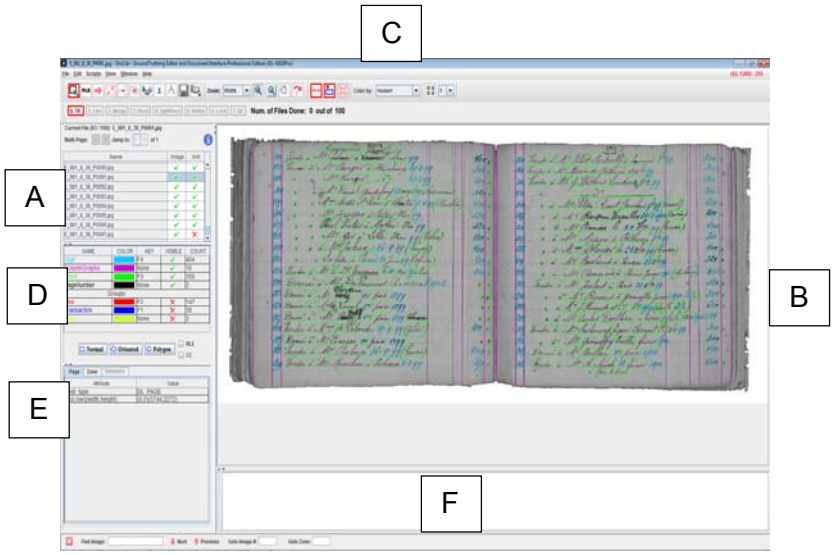


Figure 19: Interface GEDI

Interface

L'interface est implémentée en Java et fournit un ensemble d'outil pour créer, configurer et manipuler des métadonnées* correspondantes à des régions d'intérêt sur l'image.

Un seul fichier XML correspond à chaque image est affiché dans le panneau de fichier (Figure 19: A). Lorsqu'un fichier est sélectionné l'image apparait dans le panneau d'image (Figure 19 : B) accompagné de données créées précédemment. Les spécifications des données créées seront mentionnées dans la partie suivante. Un ensemble d'outils est consacré pour charger, enregistrer et manipuler l'affichage de la métadonnée (Figure 19 : C). Le type d'une zone est configurable de même pour ses attributs. L'interface prévoit un mécanisme de sélection pour choisir le type d'une zone dessinée et fournit des informations telles que le nombre des zones présentes (Figure 19: D) Pour toute page ou zone, toutes les valeurs d'un attribut sont affichées en format texte dans la fenêtre des attributs (Figure 19: E). Finalement un bar d'outils de recherche (Figure 19 : F) aide à parcourir le répertoire de travail.

**métadonnée: donnée servant à définir ou décrire une autre donnée quel que soit son support.*

Spécification GEDI

La métadonnée :

La métadonnée GEDI est une représentation basée sur XML qui représente les informations aux niveaux document, page et zone. Chaque zone a ses attributs qui peuvent être configurés à l'aide de l'interface et fournies sous la forme d'un fichier GEDIConfig.XML aux annotateurs au début d'un projet. L'interface GEDI enregistre le nom de l'utilisateur qui modifie le fichier et les dates où il l'a modifié.

```
<?xml version="1.0" encoding="UTF-8"?>
<!--GEDI was developed at Language and Media Processing Laboratory, University of Maryland.-->

<GEDI xmlns="http://lamp.cfar.umd.edu/media/projects/GEDI/" GEDI_version="2.4.2" GEDI_date="01/09/2014">
    <USER name="Dalii" date="3/23/2017 10:10" dateFormat="mm/dd/yyyy hh:mm">    </USER>
    <DL_DOCUMENT src="E_981_8_38_P0090.jpg" NrOfPages="1" docTag="xml">
        <DL_PAGE gedi_type="DL_PAGE" src="E_981_8_38_P0090.jpg" pageID="1" width="5744" height="2272">
            <DL_ZONE gedi_type="Word" id="1" col="720" row="188" width="214" height="76" Name="false" Type="Normal" Content="Donné"> </DL_ZONE>
            <DL_ZONE gedi_type="Word" id="2" col="732" row="274" width="240" height="94" Name="false" Type="Normal" Content="Vendu"> </DL_ZONE>
            <DL_ZONE gedi_type="Word" id="3" col="728" row="368" width="202" height="86" Name="false" Type="Normal" Content="Vendu"> </DL_ZONE>
            <DL_ZONE gedi_type="Word" id="4" col="710" row="456" width="178" height="86" Name="false" Type="Normal" Content="Vendu"> </DL_ZONE>
            <DL_ZONE gedi_type="Word" id="5" col="724" row="544" width="170" height="88" Name="false" Type="Normal" Content="Vendu"> </DL_ZONE>
            <DL_ZONE gedi_type="Word" id="6" col="718" row="640" width="190" height="86" Name="false" Type="Normal" Content="Vendu"> </DL_ZONE>
            <DL_ZONE gedi_type="Word" id="7" col="682" row="734" width="256" height="88" Name="false" Type="Normal" Content="Donné"> </DL_ZONE>
            <DL_ZONE gedi_type="Word" id="8" col="724" row="824" width="182" height="88" Name="false" Type="Normal" Content="Vendu"> </DL_ZONE>
            <DL_ZONE gedi_type="Word" id="9" col="708" row="918" width="242" height="86" Name="false" Type="Normal" Content="Donné"> </DL_ZONE>
            <DL_ZONE gedi_type="Word" id="10" col="734" row="1000" width="190" height="102" Name="false" Type="Normal" Content="Vendu"> </DL_ZONE>
            <DL_ZONE gedi_type="Word" id="11" col="732" row="1106" width="180" height="90" Name="false" Type="Normal" Content="Vendu"> </DL_ZONE>
            <DL_ZONE gedi_type="Word" id="12" col="720" row="1200" width="186" height="92" Name="false" Type="Normal" Content="Vendu"> </DL_ZONE>
            <DL_ZONE gedi_type="Word" id="13" col="722" row="1294" width="190" height="98" Name="false" Type="Normal" Content="Vendu"> </DL_ZONE>
            <DL_ZONE gedi_type="Word" id="14" col="716" row="1396" width="164" height="82" Name="false" Type="Normal" Content="Vendu"> </DL_ZONE>
            <DL_ZONE gedi_type="Word" id="15" col="906" row="1434" width="44" height="44" Name="false" Type="Normal" Content="à"> </DL_ZONE>
            <DL_ZONE gedi_type="Word" id="16" col="922" row="1344" width="50" height="48" Name="false" Type="Normal" Content="à"> </DL_ZONE>
            <DL_ZONE gedi_type="Word" id="17" col="922" row="1248" width="58" height="46" Name="false" Type="Normal" Content="à"> </DL_ZONE>
            <DL_ZONE gedi_type="Word" id="18" col="948" row="1148" width="54" height="50" Name="false" Type="Normal" Content="à"> </DL_ZONE>
```

Figure 20: Format de la métadonnée générée

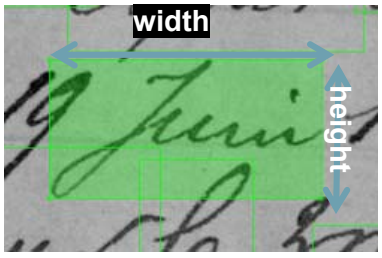
Les Zones :

Chaque zone GEDI possède un ID unique, représente un emplacement physique sur la page, et peut être représentée par une boîte de délimitation. Nous définissons différents types de zones qui peuvent avoir des différents d'attributs. Une Zone qui décrit un mot par exemple a un contenu, peut être un nom ou un mot barré. Une Zone peut être configurée pour regrouper d'autres types de zone. Une zone « Line » par exemple regroupe les zones « Word » et « Digit » qui se trouvent sur la même ligne.

Les attributs :

Chaque zone et initialement configurer par le GUI pour avoir un nom, une couleur et des valeurs par défaut. Quelques attribut sont «réservés » que chaque zone doit avoir :

- Gedi-type : a le même contenu que le nom attribué à la zone correspondante.
- (col, row)(width, height) : indique la taille et la position d'un boîtier dessiné et les coordonnées minimales. (xmin, ymin)(largeur, hauteur).



- ID : Chaque zone (boite englobante) à un ID unique que GEDI génère en fonction de l'ordre de création d'une zone.

- NextZoneID : Avec la fonctionnalité « Reading order » on peut définir un ordre pour parcourir les zones par leurs ID, cet attribut aide à implémenter cette fonctionnalité.
- Group : indique que cette zone est un regroupement d'autres zones existantes.
- Eléments : L'ensemble des ID des zones regroupées par une autre zone.

Guide d'utilisation

Ouverture d'une session

Maintenant que la configuration est prête, le fichier GEDIConfig.XML est remis aux annotateurs avec le dossier qui contient les images à annoter. Avant de se lancer, GEDI demande un nom d'utilisateur et de charger un fichier de configuration s'il existe déjà.

Run GEDI !

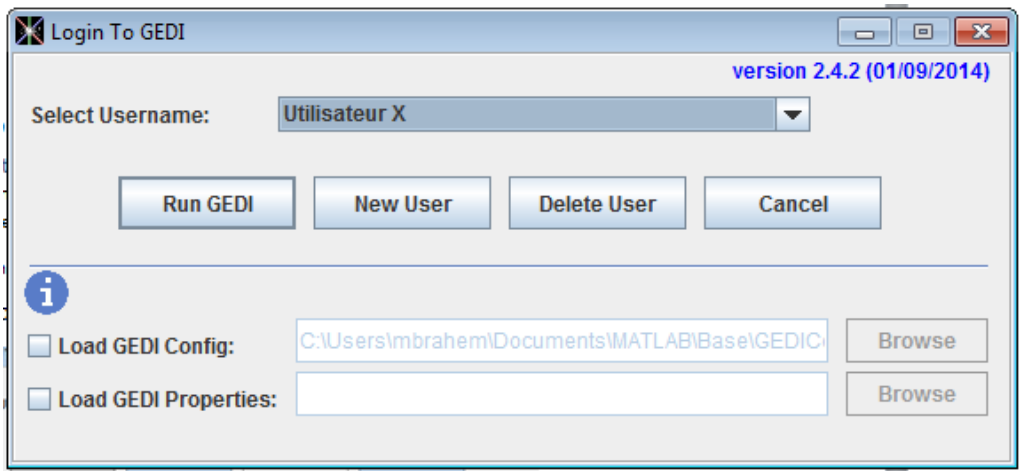
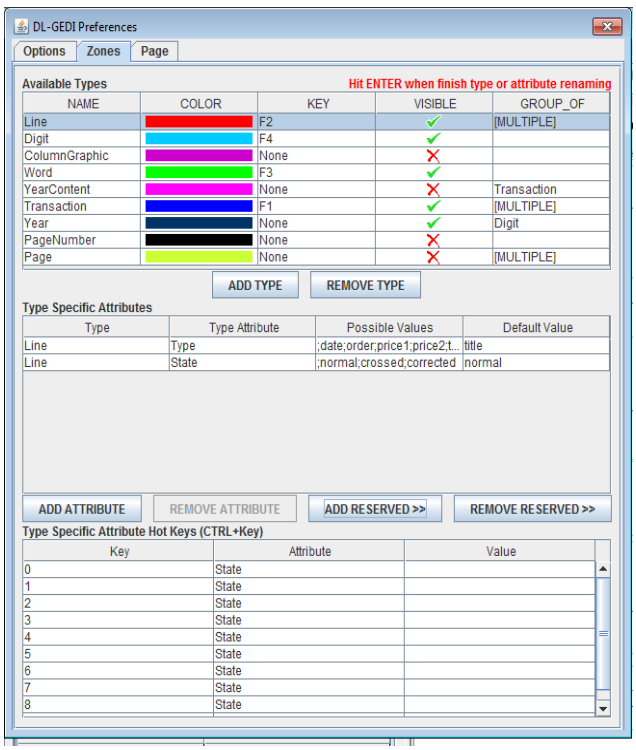


Figure 21: fenêtre de login

Configuration des Zones et des attributs :

La Configuration se fait par Edit>Préférences>Zones



- Pour ajouter une nouvelle zone, on clique sur ADD TYPE
- Une fois créée, on peut la nommer, choisir sa couleur, son raccourci clavier et sa visibilité sur l'interface.
- GROUP_OF consiste à regrouper un ou plusieurs types de zone dans une même zone.
- Pour ajouter un attribut à une zone, on utilise ADD ATTRIBUTE.
- Un attribut se caractérise par un type, des valeurs possibles et une valeur par défaut.
- L'onglet Page sert à ajouter des attributs à la zone réservée DL_PAGE.
- L'onglet option est pour les préférences générales de l'interface.

Après avoir fini la configuration, l'utilisateur confirme en appuyant sur OK. Un fichier **GEDIConfig.XML** (Figure 22) est créé et prêt à être utilisé par les annotateurs.

```
<?xml version="1.0" encoding="UTF-8"?>
<!--GEDI was developed at Language and Media Processing Laboratory, University of Maryland.-->

<GEDI xmlns="http://lamp.cfar.umd.edu/media/projects/GEDI/" GEDI_version="2.4.2" GEDI_date="01/09/2014">
    <configuration>
        <type_settings>
            <type_setting_entry gedi_type="Line" color="#ff0000" visible="false" key="F2" groupOf="Digit;Word;Year">
                <type_attribute name="Type" default="" possible_values="{;date;order;price1;price2;text}" ArbitraryVal="false" > </type_attribute>
                <type_attribute name="ColumnIndex" default="" possible_values="{;0;1;2;3;4;5}" ArbitraryVal="false" > </type_attribute>
                <type_attribute name="State" default="normal" possible_values="{;normal;crossed;corrected}" ArbitraryVal="false" > </type_attribute>
            </type_setting_entry>
```

Figure 22:Configuration de la zone Line sous forme XML

Chargement de fichier/répertoire

Une fois la session est ouverte, on choisit le répertoire de fichier adéquat File > LoadFile/Directory > Browse. Le répertoire s'affiche dans le panneau de fichier (Figure 23). Le panneau contient des informations telles que les noms des images chargées, l'état du fichier XML associé (coché vert s'il existe, croix rouge si l'image est encore vierge).

A

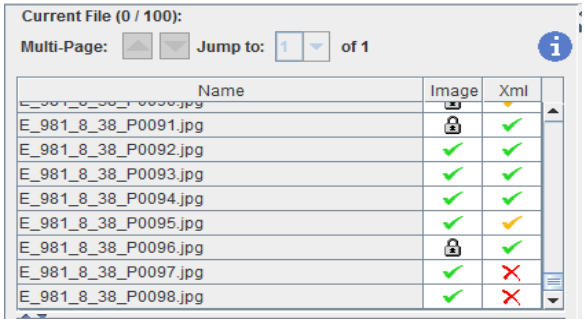


Figure 23: Panneau de fichier

Annotation des zones et ajout d'attributs

L'utilisateur sélectionne d'abord l'image qu'il souhaite annoter L'image se charge dans le panneau d'image (Figure 19 :B). À l'aide d'un panneau qui contient les configurations établies. Pour annoter une zone, on sélectionne le type de la zone qu'on désire créer, puis à l'aide du curseur, on détermine les limites de cette zone (Figure 25).

D

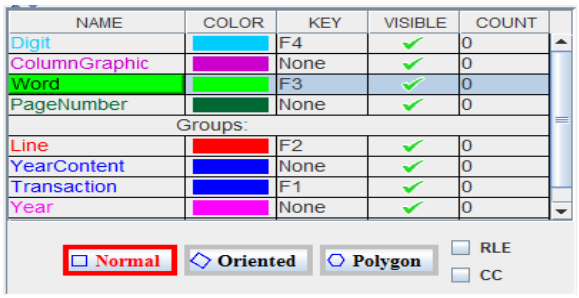


Figure 24: Panneau de configuration des zones

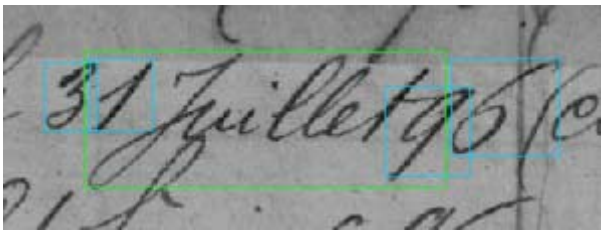


Figure 25: exemple d'un mot encadré

Pour ajouter ou modifier un attribut à une zone, il faut :

- Sélectionner la zone (figure 26)
- Utiliser la fenêtre des attributs qui s'affiche par suite (figure 9)

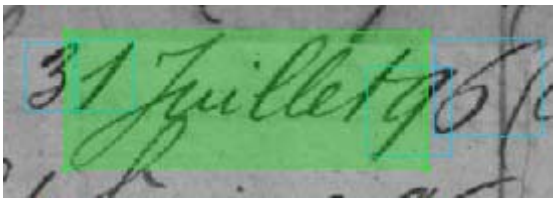


Figure 26: Zone sélectionnée

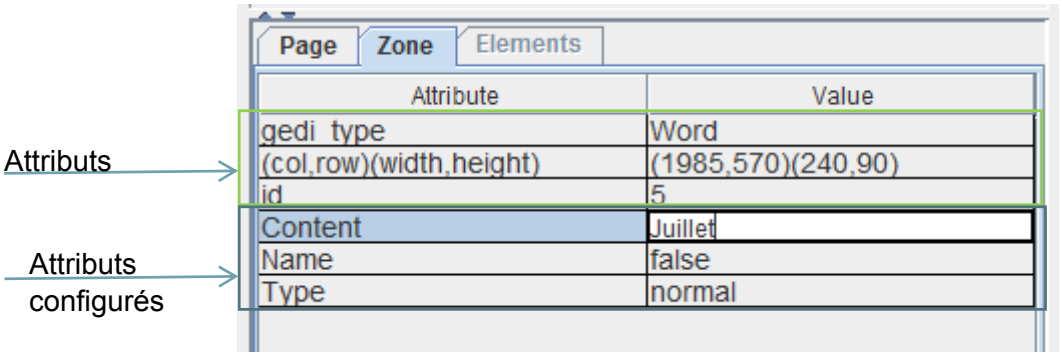


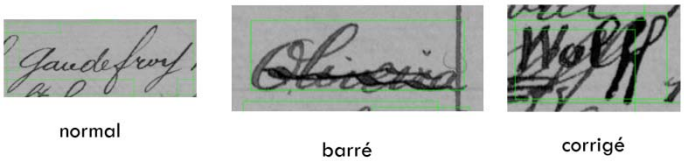
Figure 27: Les attributs d'une zone de type "Word"

Configuration utilisée

Les zones de base et leurs attributs

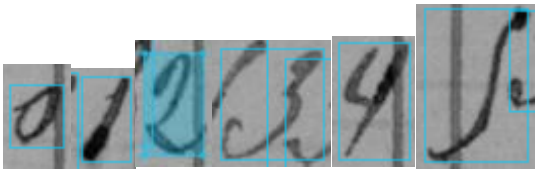
- Word :

Un mot peut être :

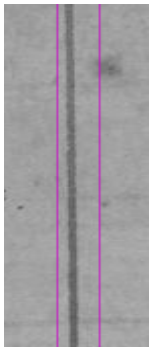


- Digit :

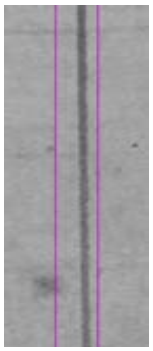
Les Chiffres de 0 à 9 sont annotés :



- ColumnGraphic :

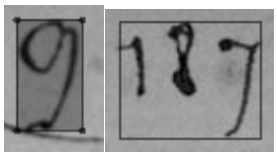


Left



Right

- PageNumber :



Regroupement des zones

Après l'étude de la structure du registre dans la première partie et afin de réaliser la structure du fichier XML désirée, les mots et les chiffres ont été regroupés par d'autres zones en utilisant la fonctionnalité « GROUP_OF » pendant la configuration des zones (voir configuration des zones et des attributs).

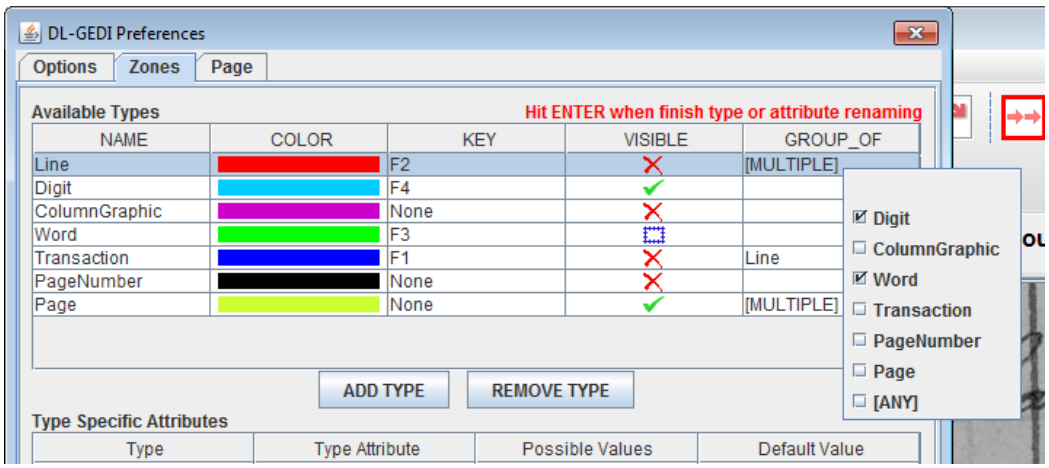


Figure 28: Line is a GROUP_OF Digit and Word

- Line :

Une zone de type “Line” est un ensemble de mots et/ou chiffres.

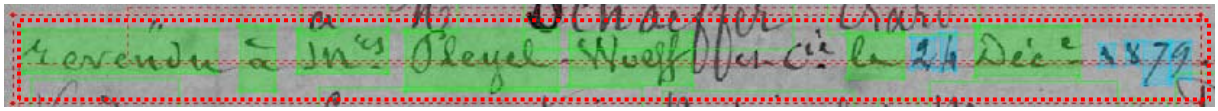


Figure 29: Ligne de type texte dans l’état normal

- Les attributs de la zone Line :
  - State : Normal – Crossed - Corrected
  - Type : Order – text – price1 – price2 – date - Title
  - Elements : indique toujours les ID des zones appartenant à la zone Line

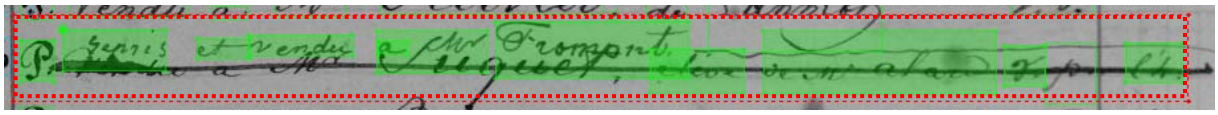


Figure 30: « Line » de type texte dans l'état « Crossed »

- Transaction :

La zone de type transaction décrit une transaction complète, elle commence généralement par une ligne de type order et finit par une ligne de type price 2. Une transaction n’admet qu’un seul numéro d’ordre.

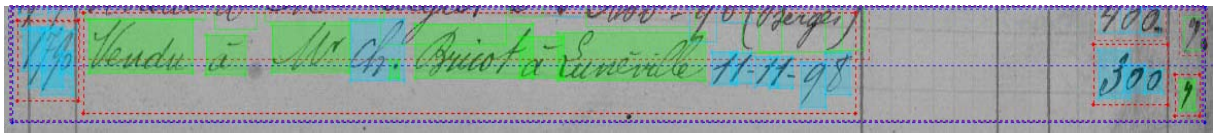


Figure 31: Transaction d'ordre 1750

- Page :
  La zone Page est constituée d’un ensemble de transactions, colonnes graphiques, numéro de la page et, parfois, une zone « Line » de type Title.

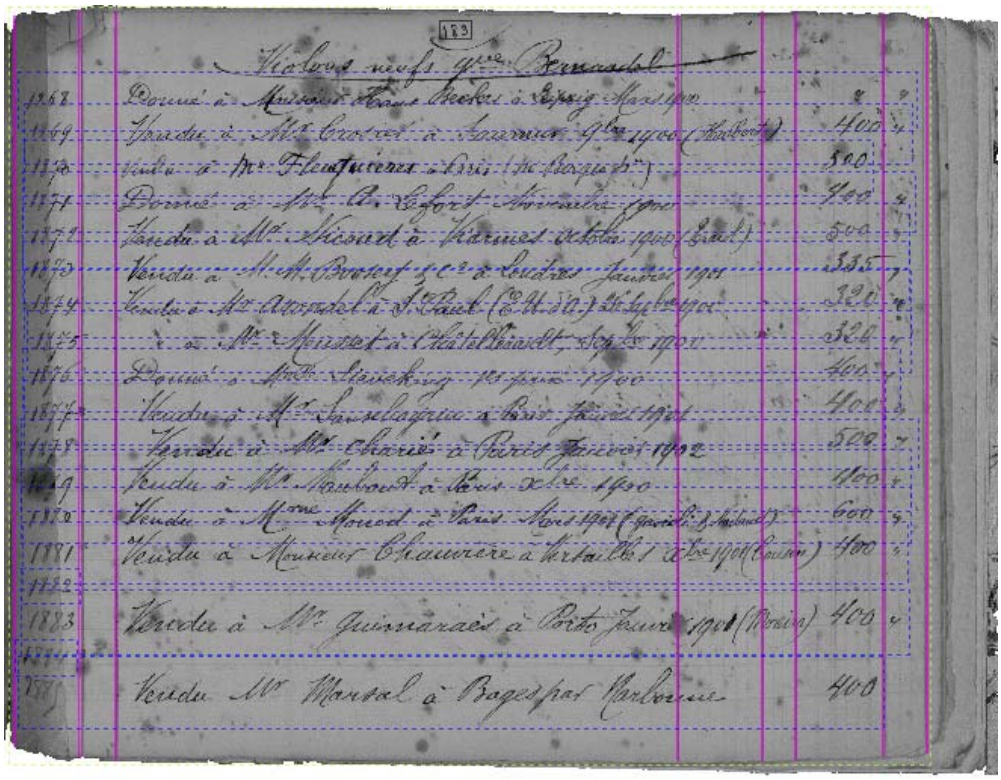
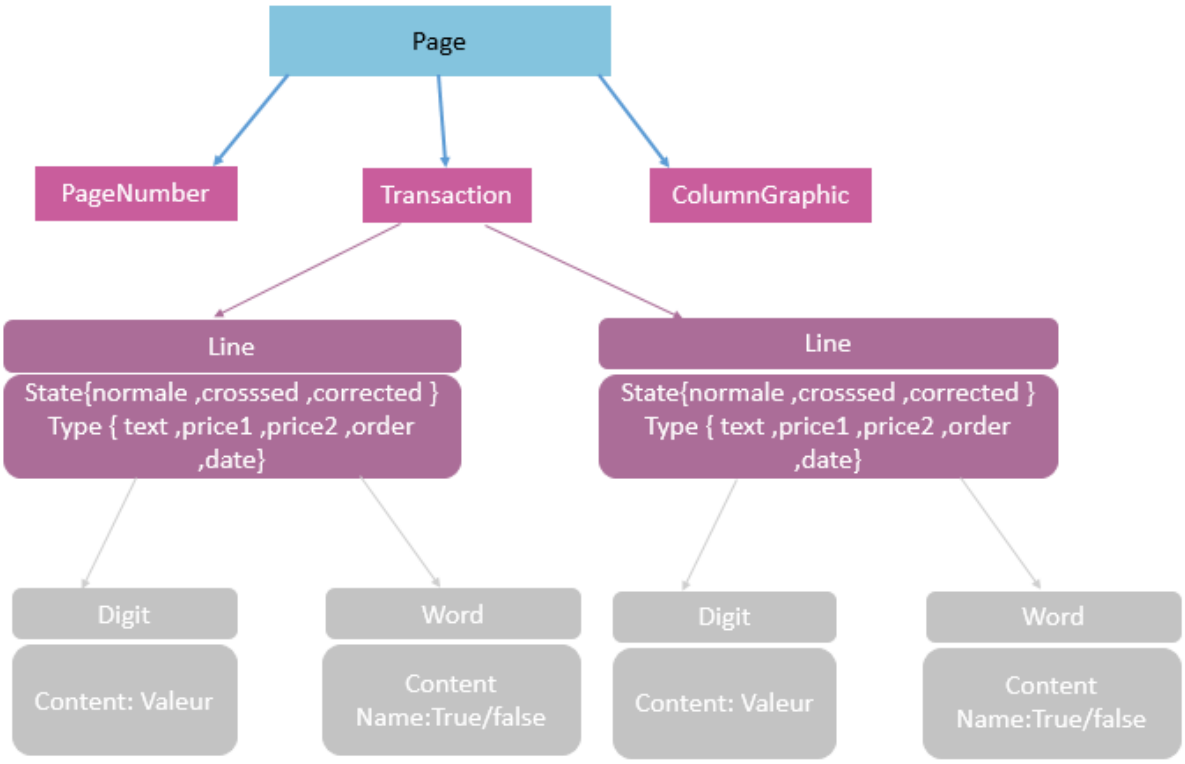


Figure 32: Exemple d'une Zone de type Page

On utilisant cette configuration on obtient la structure suivante :



Analyse de la métadonnée

La structure du fichier XML généré par GEDI est la suivante :

Le prologue

La première partie appelée prologue est une déclaration XML qui comporte le numéro de version et la déclaration d'encodage.

```
1    <?xml version="1.0" encoding="UTF-8"?>
```

La deuxième ligne est le commentaire suivant :

```
2    <!--GEDI was developed at Language and Media Processing Laboratory, University of Maryland.-->
```

L'arbre des éléments

Elle est constituée d'une hiérarchie de balises comportant éventuellement des attributs. Prenons l'exemple d'une image avec une seule zone :

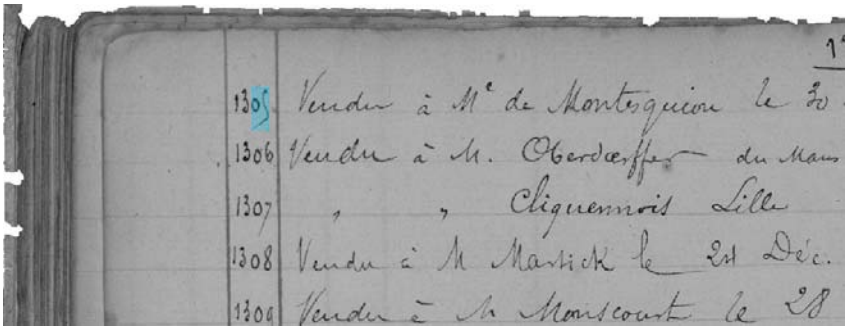


Figure 33: image annotée avec une seule zone

On obtient alors cet arbre des éléments :

```
4  <GEDI xmlns="http://lamp.cfar.umd.edu/media/projects/GEDI/" GEDI_version="2.4.2" GEDI_date="01/09/2014">
5      <USER name="UtilisateurX" date="2/11/2017 09:15" dateFormat="mm/dd/yyyy hh:mm"> </USER>
6      <DL_DOCUMENT src="E_981_8_38_P0070.jpg" NrOfPages="1" docTag="xml">
7          <DL_PAGE gedi_type="DL_PAGE" src="E_981_8_38_P0070.jpg" pageID="1" width="5744" height="2272">
8              <DL_ZONE gedi_type="Digit" id="1" col="568" row="244" width="28" height="72" Content="5"> </DL_ZONE>
9          </DL_PAGE>
10     </DL_DOCUMENT>
11 </GEDI>
```

Les informations qu'on peut extraire de ce fichier sont :

- La racine GEDI est l'élément racine, tous les éléments qui suivent sont contenus dans cet élément.
- Les attributs name, date, dateFormat de la balise USER indiquent respectivement le nom de l'utilisateur, la date de la dernière modification et le format de la date.
- DL_DOCUMENT nous informe sur le document traité, dans cet exemple l'image est « E_981_8_38_P0070.jpg », elle admet une seule page, et la métadonnée a l'extension XML
- DL_PAGE est une zone de type DL_PAGE que GEDI crée automatiquement. Cette zone est de même taille que l'image traitée comme l'indiquent les attributs width et height. Chaque DL_PAGE a un pageID et un fichier source. Entre la paire de balises DL_PAGE existent toutes les zones annotées.
- DL_ZONE : Pour chaque zone créée par l'utilisateur une balise ouvrante DL_ZONE est créée par GEDI. Les attributs de la balise DL_ZONE représentent les attributs configurés par l'utilisateur et d'autres attributs de GEDI qui indiquent la position physique de la zone dans l'image.

Les Zones « GROUP_OF »

Les zones « GROUP_OF » ont la même représentation que les autres DL_ZONE. L'attribut éléments que GEDI ajoute aide à se déplacer du niveau page jusqu'au niveau des chiffres et des mots.

L'attribut éléments de la zone page par exemple contient les ID de toutes les transactions, les colonnes graphiques et les numéros de page :

<DL_ZONE gedi_type="Page" id="1139" col="342" row="110" width="2574" height="2028" elements="447;448;449;450;451;452;453;454;456;1008;1011;1014;1017;1020;1023;1026;1029;1032;1035;1038;1041;1044;1047;1050;1053;1056;1058;1061;1131"> </DL_ZONE>

Statistiques

La base de documents traitée contient 100 images en format JPG dont 14 images ont été annotées.

- Liste des images traitées :

E_981_8_38_P0004	*E_981_8_38_P0009*	*E_981_8_38_P0093*
E_981_8_38_P0005	*E_981_8_38_P0089*	*E_981_8_38_P0094*
E_981_8_38_P0006	*E_981_8_38_P0090*	*E_981_8_38_P0095*
E_981_8_38_P0007	*E_981_8_38_P0091*	*E_981_8_38_P0096*
E_981_8_38_P0008	*E_981_8_38_P0092*	

- Liste des fichiers XML résultants :

E_981_8_38_P0004	*E_981_8_38_P0009*	*E_981_8_38_P0093*
E_981_8_38_P0005	*E_981_8_38_P0089*	*E_981_8_38_P0094*
E_981_8_38_P0006	*E_981_8_38_P0090*	*E_981_8_38_P0095*
E_981_8_38_P0007	*E_981_8_38_P0091*	*E_981_8_38_P0096*
E_981_8_38_P0008	*E_981_8_38_P0092*	*GEDIConfig*

Zone	Compte
Word	*2949*
Digit	*3849*
Line	*1581*
Transaction	*449*
GraphicColumn	*201*
Page	*28*

Conclusion

Nous avons présenté dans ce rapport les premiers pas vers la reconnaissance automatique d'information dans les documents manuscrits. Un prétraitement est nécessaire pour avoir les meilleurs documents initiaux possibles car le reste de processus (segmentation et classification) est influencés par la qualité du prétraitement. L'annotation, d'un autre côté, nous permettra de générer une base de données pour entrainer et évaluer le système de reconnaissance automatique des mots.

[i] Ra_ Cohen1, Itshak Dinstein, Jihad El-Sana, and Klara Kedem, Using Scale-Space Anisotropic Smoothing for Text Line Extraction in Historical Documents, http://www.cs.bgu.ac.il/~rafico/LineExtraction.zip, ICFHR 2014.

[ii] Nati Kligler, https://webee.technion.ac.il/labs/cgm/Computer-Graphics-Multimedia/Software/BinCode/BinCode.zip

[iii] LANGUAGE AND MEDIA PROCESSING LABORATORY, UNIVERSITY OF OF MARYLAND. HTTPS://LAMPSRV02.UMIACS.UMD.EDU/PROJDB/PROJECT.PHP?ID=53